Australian Government
Department of Defence
Defence Science and
Technology Organisation

A Logical and Probabilistic Technique for Classification and Dimensionality Reduction for Objects with Categorical Data

Mark Porter

Intelligence, Surveillance and Reconnaissance Division
Information Sciences Laboratory

DSTO-RR-0276

ABSTRACT

A supervised learning technique, the Attribute Importance Measure (AIM) method, is proposed for the classification of objects with categorical attributes. The advantage of this method over existing techniques is its ability to perform classification and dimensionality reduction, or feature selection, with the same algorithm. The method uses probabilistic measures alongside logical concepts of sufficiency, necessity and irrelevance in providing corresponding weights to values in attribute value pairs. Finally an efficient search algorithm is developed which generates decision rules for classification. The performance of the new method is demonstrated on a commonly used machine learning data set.

Published by

DSTO Information Sciences Laboratory
PO Box 1500
Edinburgh South Australia 5111 Australia

Telephone: (08) 8259 5555
Fax: (08) 8259 6567

AR- 013-127
June 2004

APPROVED FOR PUBLIC RELEASE

A Logical and Probabilistic Technique for Classification and Dimensionality Reduction for Objects with Categorical Data

Executive Summary

The paper presents a novel method for classification, dimensionality reduction and rule discovery for data with categorical attributes. These three areas of interest in data mining are often conducted using different algorithms, while the new Attribute Importance Measure technique presented here is able to conduct all of these operations in the same algorithm. The AIM method uses a probabilistic approach, similar to that of the Naïve Bayes algorithm, with additional logical concepts of sufficiency and necessity. The goal of this paper is to present the new method and demonstrate its application, rather than to perform an extensive comparison with existing techniques.

Authors

Dr Mark Porter
Intelligence, Surveillance and Reconnaissance Division

Dr Mark Porter has a doctorate from The University of Oxford and a First Class Honours degree in Engineeing from The University of Adelaide. He commenced working at DSTO in April 2002, where his focus was initially on database research. Later his interests moved to the area of data mining and machine learning.

Contents

Notation

S	number of classes
ϖ_j	j^{th} class
N_A	number of attributes of each object
X	an object
x_k	an attribute of an object
n_k	number distinct values for the attribute x_k
x_{ki}	a value of the k^{th} attribute
$M_S(x_{ki}, \varpi_j)$	measure of sufficiency for x_{ki} being in class ϖ_j
$M_N(x_{ki}, \varpi_j)$	measure of necessity for class ϖ_j, given x_{ki}
$M_V(x_{ki})$	measure of usefulness of attribute value x_{ki} (across all S classes)
$M_A(x_k)$	measure of usefulness of attribute x_k (for all attribute values in x_k and across all S classes)
$\hat{x}_g$	vector of g attribute values. Eg. $\hat{x}_3 = \{x_{11}, x_{25}, x_{32}\}$

1. Introduction

This paper presents a new algorithm for dealing with the supervised learning problem of classifying objects with categorical, or nominal, data. It is also concerned with the problems of dimensionality reduction, or feature selection, and rule generation for categorical data. The motivation for this work is a large dataset currently being investigated in ISRD, most of whose fields contain such categorical data.

Analysis of categorical data restricts the number of existing classification algorithms that may be used. Many parametric methods (James, 1985), rely on continuous data and are inappropriate. This paper addresses pure categorical, or nominal data, not multinominal data, which is simply binned segments of the continuous scale. A common approach for such data is to code each categorical value as a dummy variable. That is, if there is a variable "colour" whose values are "red", "green" or "blue", then the following new variables might be created: "red"={0,1}, "green"={0,1} and "blue"={0,1}. This approach is useful because the new variables can now be taken as quasi-continuous, and therefore used in traditional machine learning algorithms. However, for cases, such as the one considered in ISRD, where the dataset is very large, and the number of categorical values exceedingly high, such an approach becomes unworkable as the dimensionality becomes large.

There are algorithms which accommodate categorical data without the need for the recoding described above, some of which are discussed here. Nearest-neighbour classification algorithms (k-NN) (Duda *et al.*, 2001) can perform classification of a data point based on the training set, but is unable to generate decision rules. Dimensionality reduction is also possible with this algorithms, but it is conducted through an iterative process whereby a subset of features is selected, and the accuracy of classification using those features is used as a metric for comparing different subsets of features. This type of procedure is more computationally intensive than single step solutions, such as the use of the Gini, entropy or Chi-squared measures.

Such measures are used in decision tree methods (Hastie *et al.*, 2001) which are also amenable to categorical data. There are many decision tree algorithms in the literature, for example Quinlan (1986), Lim (2000) and Frank (1998), and they are easily applicable to the dataset being considered in ISRD. However, there is an amount of inflexibility in rule generation using these methods. When descending a decision tree, from the root to a leaf, each node is positioned based on some metric, such as the Gini measure, as given above. This provides the implicit feature selection of the algorithm. Decision rules are then formed by amalgamating decisions at each branch through to a leaf node. However, variables used in splits close to the root are based on the usefulness of the variable over all classes, not each individual one. It is quite probable that the decision rules for each class may not rely on the same variables, and therefore the decision rules generated through decision trees are somewhat limited.

Probabilistic approaches are also suited to classification with categorical data. The classical probabilistic approach uses the Naïve Bayes method (Lewis, 1998 and Mosteller & Wallace, 1984), and it is this method which is the most similar to the new algorithm presented in this paper. Therefore, classification using the Naïve Bayes approach is described further below.

Consider the problem of classifying an object, X, into one of S mutually exclusive, exhaustive classes, $(\varpi_1,...,\varpi_j,...,\varpi_S)$. There are series of attributes of X, $(x_1,...,x_k,...,x_{N_A})$, whose values are all categorical. The traditional approach would be to classify X into the class where the conditional probability $P(\varpi_j \mid X)$ is maximised (Hand, 1981, p.5). Estimates of probabilities are calculated from the training set. In many applications, for example textual analysis (Sebastiani, 2002), the vector of attributes associated with X can be large, causing this conditional probability to be non-trivial to calculate. Therefore, Bayes' Theorem (Lewis, 1998) is used to put this probability in a more convenient form.

$$P(\varpi_j \mid X) = \frac{P(X \mid \varpi_j)P(\varpi_j)}{P(X)} \tag{1}$$

Maximising $P(\varpi_j \mid X)$ is effective for finding the class to which X most likely belongs. Exclusion from a class can also be deduced as $P(\varpi_j \mid X)$ approaches zero. The point of changeover between likely inclusion and likely exclusion of an object from a class is rarely mentioned in literature for the Bayesian approach. James (1985, pp. 70-72) proposes a limit of $(1-t)$ to $P(\varpi_j \mid X)$, below which an object is excluded from classification in ϖ_j. The parameter t here is arbitrarily chosen by the user, rather than being a rigorous criterion. The method presented in this paper makes use of a more rigorously defined changeover point, and it will be shown that it has a logical significance which, when combined with other logical conditions, leads to further functionality of the classification algorithm.

2. Attribute value distributions and logical conditions

For brevity and clarity, let us consider just one on the attributes of X, x_k say, which has categorical attribute values $\{x_{k1},...,x_{ki},...,x_{kn_k}\}$, where n_k is the number of distinct categorical data values for attribute x_k. Figure 1 gives an example contingency table showing the distribution of these attribute values amongst three classes for 100 training objects.

	x_{k1}	x_{k2}	x_{k3}
ϖ_1	15	0	0
ϖ_2	5	30	0
ϖ_3	20	7	23

Figure 1 Example data for $n_k = 3$ *and* S = 3.

Firstly, we can see from the figure that all of the occurrences of x_{k3} fall in class ϖ_3. Therefore, if an object has the attribute value x_{k3}, that is *sufficient* information to classify the object into class ϖ_3. Equally, given that no objects with attribute value x_{k3} fall into ϖ_1 or ϖ_2, it is *sufficient* to know that an object has attribute value x_{k3} to know that it *does not* fall into either of classes ϖ_1 or ϖ_2. This example shows that these conditions of perfect sufficiency occur when a column in the contingency table of Figure 1 is non-zero in just one cell. This can be summarised as follows.

$$\text{Sufficient for inclusion in } \varpi_j \text{ if} \qquad P(\varpi_j \mid x_{ki}) = 1 \tag{2}$$

$$\text{Sufficient for exclusion from } \varpi_j \text{ if} \qquad P(\varpi_j \mid x_{ki}) = 0 \tag{3}$$

Consider now the attribute values which occur for class ϖ_1. Figure 1 shows that only x_{k1} falls in ϖ_1, but that objects with attribute value x_{k1} can also be found in other classes. Therefore, it is *necessary* for an object to have attribute value x_{k1} to be in class ϖ_1, but not sufficient. The negative necessity condition is exemplified by x_{k3} and ϖ_1 or ϖ_2, where it is *necessary* for an object to have x_{k3} for it *not* to be in either class. These cases show that perfect conditions of necessity result from rows in the contingency table of Figure 1 containing only one non-zero element. More specifically,

$$\text{Necessary for inclusion in } \varpi_j \text{ if} \qquad P(x_{ki} \mid \varpi_j) = 1 \tag{4}$$

$$\text{Necessary for exclusion from } \varpi_j \text{ if} \qquad P(x_{ki} \mid \varpi_j) = 0 \tag{5}$$

Note that Equations 3 and 5 are logically equivalent. This implies that a condition of negative sufficiency is the same as a condition of negative necessity. That is, if it is necessary that an object *not* have attribute value x_{ki} to be in class ϖ_j, then this is sufficient information to know that an object that has x_{ki} is not in ϖ_j.

Note that the logical conditions of sufficiency and necessity are not mutually exclusive. The combined condition occurs when a cell contains the only non-zero value in its row and column. In this case there is direct mapping between the attribute value and the class.

Table 1 Logical conditions and their probabilistic criteria

Logical condition	Direction	Criteria	$M_S(x_{ki},\varpi_j)$	$M_N(x_{ki},\varpi_j)$
Sufficiency	Inclusion	$P(x_{ki} \mid \varpi_j) = \frac{P(x_{ki})}{P(\varpi_j)}$	1	N/A
Sufficiency or Necessity	Exclusion	$P(x_{ki} \mid \varpi_j) = 0$	-1	-1
Necessity	Inclusion	$P(x_{ki} \mid \varpi_j) = 1$	N/A	1
Irrelevance	N/A	$P(x_{ki} \mid \varpi_j) = P(x_{ki})$	0	0

The final logical condition introduced here is that of irrelevance. If $P(\varpi_j \mid x_{ki}) = P(\varpi_j)$, then the condition that an object has attribute value x_{ki} does not affect the classification of the object into class ϖ_j. This is the definition that the occurrence of x_{ki} and the classification of an object into ϖ_j are independent events. Gennari *et al.* (1989, section 5.5) has previously expressed irrelevance in these terms. In Figure 1, $P(\varpi_3 \mid x_{k1}) = P(\varpi_3) = 0.5$, signifying that the presence of x_{k1} is *irrelevant* to classification into ϖ_3. Note, however, that this conditional probability is the maximum for x_{k1} and any of the classes, and would therefore have been the optimal classification mechanism under a traditional Naïve Bayes scheme. This is similar to the Bayesian constant rule problem (James, 1985, pp. 159-160) where every case is assigned to the class which occurs most often. James (1985) notes that this is a particular problem for categorical data. By defining a point of irrelevance the new method avoids this problem.

Bayes' Theorem is used to express the criteria above in terms of $P(x_{ki} \mid \varpi_j)$. The resulting expressions are presented in Table 1. The rightmost two columns of the table will be discussed in the next section.

3. Definition of Attribute Importance Measures

3.1 Sufficiency and necessity measures

It is proposed that two measures be introduced: $M_S(x_{ki},\varpi_j)$, which gives a measure of the sufficiency that an object be in class ϖ_j given attribute value x_{ki}; and, $M_N(x_{ki},\varpi_j)$, which gives a measure of the necessity that an object have attribute value x_{ki} to be in class ϖ_j. The measures attain their maxima of ± 1 when the perfect logical conditions of Table 1 are achieved, and are zero for the condition of irrelevance (as presented in Table 1). For

simplicity, linear interpolation is used between these points, as shown in Figure 2, resulting in Equations 6 to 8 below. Note that this derivation is in terms of a single attribute value: the method will be applied to vectors of attribute values later in the paper.

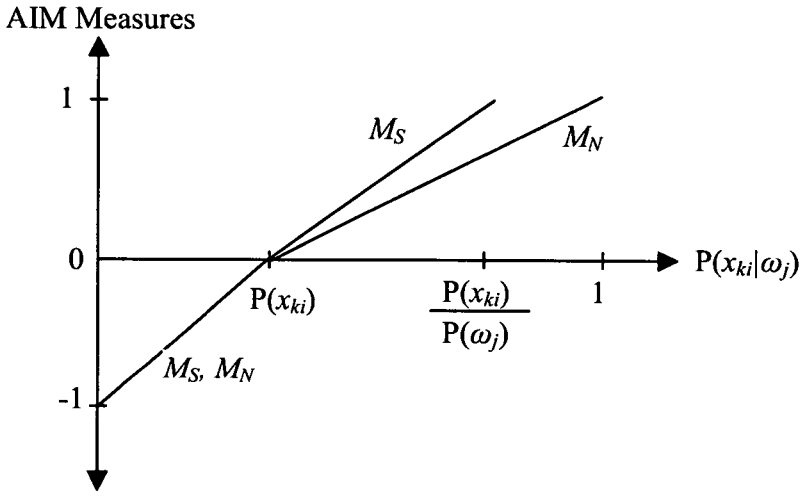

Figure 2 $M_S(x_{ki},\varpi_j)$ *and* $M_N(x_{ki},\varpi_j)$ *versus* $P(x_{ki} \mid \varpi_j)$

$$M_S(x_{ki},\varpi_j) = \frac{P(x_{ki} \mid \varpi_j) - P(x_{ki})}{\frac{P(x_{ki})}{P(\varpi_j)} - P(x_{ki})} \qquad \text{if } P(x_{ki} \mid \varpi_j) > P(x_{ki}) \quad (6)$$

$$M_N(x_{ki},\varpi_j) = \frac{P(x_{ki} \mid \varpi_j) - P(x_{ki})}{1 - P(x_{ki})} \qquad \text{if } P(x_{ki} \mid \varpi_j) > P(x_{ki}) \quad (7)$$

$$M_S(x_{ki},\varpi_j) = M_N(x_{ki},\varpi_j) = \frac{P(x_{ki} \mid \varpi_j) - P(x_{ki})}{P(x_{ki})} \qquad \text{if } P(x_{ki} \mid \varpi_j) \le P(x_{ki}) \quad (8)$$

It can easily be shown that a perfect necessity condition is unachievable if $P(x_{ki}) < P(\varpi_j)$, while no sufficiency condition can be obtained if $P(x_{ki}) > P(\varpi_j)$.

Having calculated these measures, an object may be classified into a class where the sufficiency measure is highest (and positive). This is analogous to Bayesian classification where the conditional probability of Equation 1 is at its maximum.

3.2 The Attribute Value Measure

Equations 6 to 8 give measures of the usefulness of attribute values in classifying an object into a particular class. A simple summation scheme can now reveal the classifying power of an attribute value x_{ki} over all classes. A summation is logical as the zeros for the AIM measures, M_S and M_N, occur when the attribute value is irrelevant to classification. This contrasts to traditional approach where $P(x_{ki} \mid w_j) = 0$ in fact relates to a case of perfect prediction that an object not belong to class ϖ_j. The measure of the predictive power of an attribute value, $M_V(x_{ki})$, in general terms is given by

$$M_V(x_{ki}) = \sum_{j=1}^{S} F\big(M_S(x_{ki},\varpi_j), M_N(x_{ki},\varpi_j)\big)W_j \tag{9}$$

where F is some function of the measures of sufficiency and necessity, and W_j is a weight parameter. An obvious choice for W_j is $P(\varpi_j)$. A convenient form of the function F is a linear combination of the square of the measures, which results in $M_V(x_{ki})$ being zero when all the individual measures are zero (that is for total irrelevance), and achieving a maximum of one when the individual measures are all at their maxima (either positive or negative). Such a function is given in Equation 9a.

$$M_V(x_{ki}) = \sum_{j=1}^{S} \frac{1}{2}\big(M_S(x_{ki},\varpi_j)^2 + M_N(x_{ki},\varpi_j)^2\big)P(\varpi_j) \tag{9a}$$

A measure of the usefulness of an attribute value to classification is dependent on *how* that attribute value is to be used in classification. The measure presented in Equation 9a gives equal weighting to the measures of sufficiency and necessity. In many applications, however, it is only the sufficiency measure that will be of use, as it is sufficiency which maps a given attribute value directly to a class. Therefore, a second variation for the function F of Equation 9 is presented below for use in applications where the condition of necessity is not used in classification. Note that the modified measure maintains its range of [0,1].

$$M_V(x_{ki}) = \sum_{j=1}^{S} M_S(x_{ki},\varpi_j)^2 P(\varpi_j) \tag{9b}$$

3.3 The Attribute Measure

Just as $M_V(x_{ki})$ provides a measure of the predictive power of attribute value x_{ki}, a summation scheme over all an attribute's values can provide a measure of the classification power contained in the attribute as a whole, x_k. For example, if the

individual values of an attribute all have $M_V(x_{ki}) = 0$, then none of them provide useful information for classification and therefore the attribute x_k as a whole has no predictive power. The measure for an attribute as a whole, $M_A(x_k)$, is given in general terms by

$$M_A(x_k) = \sum_{i=1}^{n_k} M_V(x_{ki}) W_i \tag{10}$$

where W_i is a weight parameter. A natural choice for this weight is $P(x_{ki})$, giving

$$M_A(x_k) = \sum_{i=1}^{n_k} M_V(x_{ki}) P(x_{ki}) \tag{10a}$$

The bounds for $M_A(x_k)$ from Equation 10a are [0,1]. The minimum and maximum of $M_A(x_k)$ are achieved when the sufficiency and necessity measures for each attribute value in each class are $M_S(x_{ki}, \varpi_j) = M_N(x_{ki}, \varpi_j) = 0$ and $M_S(x_{ki}, \varpi_j) = M_N(x_{ki}, \varpi_j) = \pm 1$ respectively. Note that if the attribute value measures, $M_V(x_{ki})$, are calculated omitting the necessity condition, as in Equation 9a, then the corresponding attribute measure, $M_A(x_k)$, will also be a measure based only on sufficiency results. In this case only the sufficiency measures need be maximal for the attribute measure to achieve a value of one.

Dimensionality reduction can be undertaken based on a ranking of attributes according to the attribute measure $M_A(x_k)$.

4. The Attribute Measure and the test for independence

In the preceding section, the AIM measure for the attribute as a whole, M_A, was interpreted as being a measure of how useful the attribute is to classification. The ability to perfectly classify, based on just one attribute, is achieved when $M_A = 1$. When $M_A \approx 0$, each attribute value is assumed to be irrelevant to classification. This state corresponds to independence between the attribute values and the classes, and therefore may be compared with traditional tests for independence, such as the chi squared (χ^2) test (Anderson *et al.*, 1994).

It can be shown that a chi squared statistic for analysing a contingency table such as Figure 1 is

$$\frac{\chi^2}{N} = \sum_{i=1}^{n_k} \sum_{j=1}^{S} \left(\frac{P(x_{ki} \mid \varpi_j) - P(x_{ki})}{P(x_{ki})} \right)^2 P(\varpi_j) P(x_{ki}) \qquad (11)$$

where N is the number of observations and the probabilities are estimated from the contingency table.

Now consider M_A using the form of M_V from Equation 9b. Substituting Equations 6, 8 and 9b into Equation 10a gives

$$M_A(x_k) = \sum_{i=1}^{n_k} \sum_{j=1}^{S} \left(\frac{P(x_{ki} \mid \varpi_j) - P(x_{ki})}{D} \right)^2 P(\varpi_j) P(x_{ki}) \qquad (12)$$

where D is the denominator of Equations 6 and 8. The similarities between Equations 11 and 12 are obvious, and indeed in the cases where $P(x_{ki} \mid \varpi_j) < P(x_{ki})$ the terms within the summations are identical.

It is known that the asymptotic distribution of the chi squared statistic is independent of the distribution of the classes and attribute values. Further, it is dependent only on the degrees of freedom of the system under investigation, which in this case is the sum of the number of rows and columns in the contingency table less two. A simulation study was undertaken to demonstrate that these properties also apply to the M_A measure.

4.1 Simulation study of M_A under independence

In the simulation study, a contingency table was filled according to predefined probability distribution functions. These functions, nominally labelled "uniform", "quadratic" and "peak", are shown in Figure 3 for a 10 by 10 table, and were applied independently to the distributions of the classes and attribute values. That is, the attribute values and classes were independent of each other. The N objects in the contingency table were therefore independently identically distributed (i.i.d.). With the table filled, the M_A measure was calculated (using Equation 12) and recorded. This procedure was repeated 10,000 times to give a distribution of the M_A measure. This whole process was in turn repeated a number of times, varying the probability distribution functions of the classes and attribute values each time.

Table 2 gives the details of each case considered in the simulation study for a 10 by 10 contingency table (18 degrees of freedom). The results are plotted in Figure 4 in terms of $M_A N$ to be dimensionally consistent with chi squared, where N is the number of objects classified. The similarity of the histograms in the figure suggests the populations for each case are identical. Using the Mann-Whitney-Wilcoxon test (Anderson *et al.*, 1994, pp. 729-734), the hypothesis that each of the populations are identical was retained with a 5%

significance level in all cases. It can therefore be said that $M_A N$ is independent of the distribution of attribute values and classes in the case where these are independent of each other. This is also a property of the chi squared distribution.

Table 2 Details of the cases used in the simulations

Case	N	Attribute Value pdf	Class pdf
1	10,000	Uniform	Uniform
2	10,000	Peak	Peak
3	10,000	Quadratic	Quadratic
4	10,000	Quadratic	Uniform
5	5,000	Quadratic	Quadratic

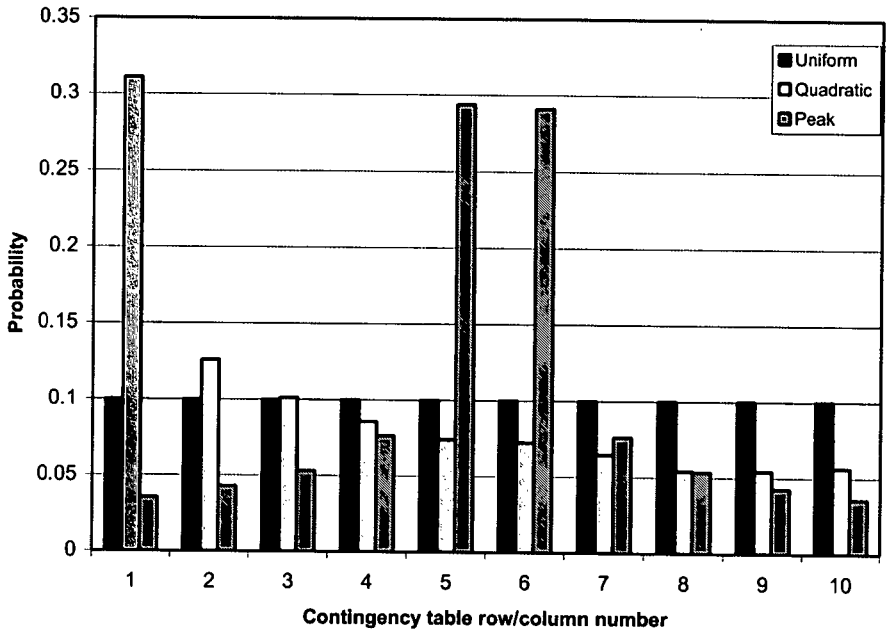

Figure 3 Probability density functions used in stochastic analysis

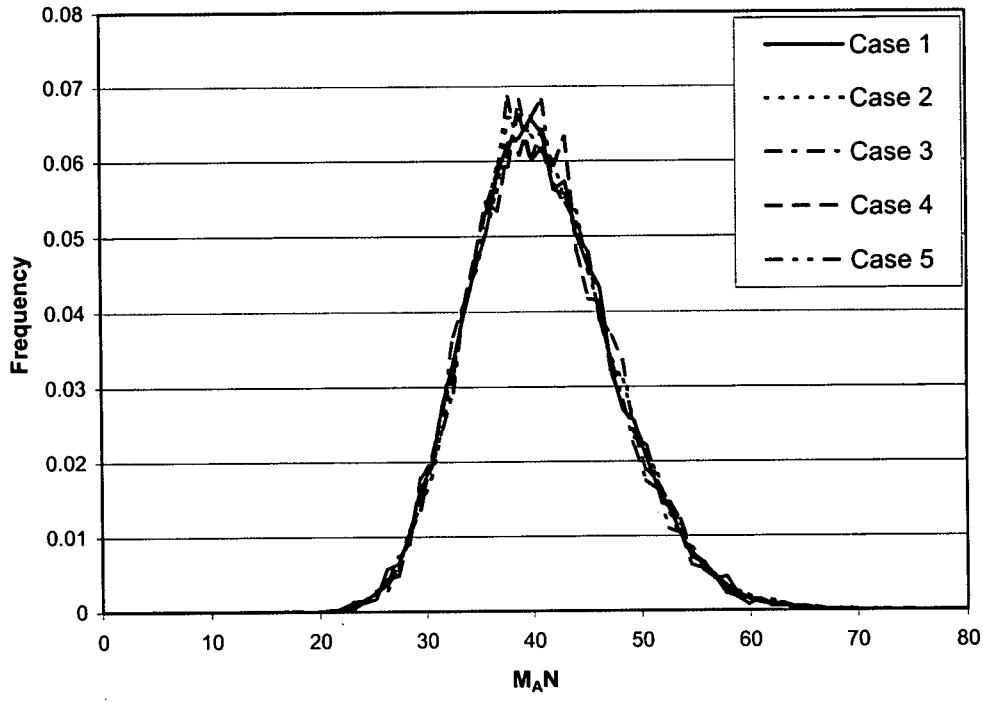

Figure 4 Histogram of $M_A N$ for the case of independence

Further study used differently sized contingency tables, which therefore had different degrees of freedom. It was found that the distribution of $M_A N$ varied with the degrees of freedom. This dependency of the distribution only on the degrees of freedom is also a property of chi squared, indicating the similarity between the two statistics.

5. Application to vectors and expressions

The derivation of the measures of necessity and sufficiency in Section 3.1 was expressed in terms of a single attribute value, x_{ki}. The measures for vectors of attribute values, such as $M_S\left((x_{19}, x_{21}, x_{35}), \varpi_j\right)$ and $M_N\left((x_{19}, x_{21}, x_{35}), \varpi_j\right)$, or expressions, such as $M_S\left((x_{19} \cup x_{21}), \varpi_j\right)$, are now considered. They rely on substitution of $P(x_{ki})$ and $P(x_{ki} \mid \varpi_j)$ in Equations 6 to 8 with corresponding expressions for the combined attribute value vector or expression. In the case of a vector of attribute values, $\hat{x}_g = (x_{1i}, \ldots, x_{ki}, \ldots, x_{gi})$, the substituted expressions are $P(\hat{x}_g)$ and $P(\hat{x}_g \mid \varpi_j)$. The assumptions of dependence or independence made when calculating these values have repercussions for the AIM method, and are therefore addressed below. Note that, for

consistency, $P(\hat{x}_g)$ and $P(\hat{x}_g \mid \varpi_j)$ should both be calculated under the same assumption, whether this be independence or dependence.

5.1 Independent attributes

If the occurrence of values between attributes is assumed to be independent, then the following is true.

$$P(\hat{x}_g) = P\left((x_{1i},...,x_{ki},...,x_{gi})\right) = \prod_{k=1}^{g} P(x_{ki}) \tag{13}$$

Further, the conditional probability for a vector of attribute values and a class is commonly decomposed thus (Lewis, 1998):

$$P(\hat{x}_g \mid \varpi_j) = P\left((x_{1i},...,x_{ki},...,x_{gi}) \mid \varpi_j\right) = \prod_{k=1}^{g} P(x_{ki} \mid \varpi_j) \tag{14}$$

The assumption of independence is useful where there are insufficient samples to adequately calculate the probabilities for vectors directly, but where the ability to compose the probabilities, as in Equations 13 and 14, leads to acceptable approximations. Note that in cases where independence is assumed, but is not reasonable, the sufficiency and necessity measures may exceed the theoretical bounds of [-1,1].

5.2 Dependent attributes

The assumption of dependent attribute values is a more rigorous approach and leads to more accurate sufficiency and necessity measures, but may be more difficult to calculate. In particular, where the vector length is long, the particular combination of attribute values being considered may not have occurred in the training set, thus inhibiting accurate estimation of the probability.

6. A decision rule search algorithm

A vector of attribute values with a high sufficiency measure for a class can be thought of as a decision rule. For example, if $M_S\left((x_{11},x_{27},x_{35},x_{42}),\varpi_1\right)=1$, then the rule $(x_{11},x_{27},x_{35},x_{42}) \rightarrow \varpi_1$ can be deduced. Note that the size of the vector need not be constant between decision rules: one class may be defined by a single attribute value, while another may need information from all of the available attributes before a decision rule may be formulated. Since there are potentially many attributes, each with many values, the total number of possible vectors of attribute values for decision rules can be

large. Therefore, an efficient method by which these decision rules may be identified is of interest in order to obviate the need to investigate each possible attribute value vector.

Consider a vector of dependent attribute values which has a high sufficiency measure for a given class, say $M_S((x_{11}, x_{27}, x_{35}, x_{42}), \varpi_1) = 1$. In this case any object with the attribute value vector given can be confidently classified into set ϖ_1. If the attribute values are analysed individually, it will be found that none of them has a high sufficiency measure for class ϖ_1. However, class ϖ_1 only occurs when each of these attribute values is present. That is, it is necessary for an object to have each of x_{11}, x_{27}, x_{35} and x_{42} to be in ϖ_1. Therefore, each attribute value will individually have a high necessity measure for ϖ_1. This demonstrates that amalgamating dependent attributes values with high necessity measures leads to vectors of attribute values with a higher sufficiency measure than those of the individual attribute values. Implicit in this statement is that attribute values with negative necessity measures for a class will not lead to a higher sufficiency measure when amalgamated with other attribute values. These principles are now used to develop an algorithm which searches for the combinations of attribute values which are best at classifying objects, that is, decision rules.

The algorithm presented here iterates through the attribute values, amalgamating those with high necessity measures until a high sufficiency measure is obtained for each class. Those combinations which have a negative necessity measure are eliminated at each stage, thus reducing the number of vectors that need to be assessed. The algorithm is set out below.

DEFINE the lower limit for the sufficiency measure, M'_S, above which objects are accepted for classification. This will be dependent on the acceptable misclassification rate in particular applications, and is determined by the user.

DO WHILE $j \leq S$ (cycle through the classes)

DEFINE vectors of attribute values, $\hat{x}_g$. The initial size of each vector, $|\hat{x}_g|$, is 1, that is each vector corresponds to a single attribute value. The subscript g corresponds to each unique vector of the attribute values.

DO WHILE there are unclassified vectors.

CALCULATE $M_S(\hat{x}_g, w_j)$ and $M_N(\hat{x}_g, w_j)$ for each vector

IF $M_S(\hat{x}_g, \varpi_j) \geq M'_S$ THEN

CLASSIFY vector into set ϖ_j

REMOVE vector from further calculations for ϖ_j

END IF

REMOVE vectors where $M_N(\hat{x}_g, \varpi_j) \leq 0$ from further calculations for ϖ_j

FORM new vector combinations, using the attribute values in the remaining vectors, to increment the vector length, $|\hat{x}_g|$, by 1. The new vectors cannot contain subsets which are vectors that have been eliminated in previous iterations.

LOOP

FOR each $\hat{x}_g$ in the set of vectors for class ϖ_j

IF $M_S(\hat{x}_g, w_j) = M_N(\hat{x}_g, w_j) = 1$ THEN

REMOVE all other vectors for ϖ_j, as they are redundant due to $\hat{x}_g$ being both necessary and sufficient.

END IF

LOOP

LOOP

7. Special case of binary classification

In many applications the number of classes is just two, $S = 2$. For example, when classifying an object as being Good or Bad, Acceptable or Unacceptable, or as in Mosteller & Wallace (1984) when investigating disputed authorship between two writers. In these cases, the AIM method has some useful properties, which can simplify analysis.

If we label the two classes ϖ_1 and ϖ_2, then the following is true.

$$P(\varpi_1) = 1 - P(\varpi_2) \tag{15}$$

and

$$P(\varpi_1 \mid x_{ki}) = 1 - P(\varpi_2 \mid x_{ki}) \tag{16}$$

It can therefore be shown that

$$M_S(x_{ki}, \varpi_1) = -M_S(x_{ki}, \varpi_2) \tag{17}$$

Equations 8 and 17 show that $M_S(x_{ki}, \varpi_1)$, $M_S(x_{ki}, \varpi_2)$ and one of $M_N(x_{ki}, \varpi_1)$ or $M_N(x_{ki}, \varpi_2)$ can be expressed in terms of $M_S(x_{ki}, \varpi_1)$.

To reduce the dimension of the problem, one of the classes is associated with positive AIM measures, while the alternative class corresponds to negative sufficiency and necessity measures. For example, the set of "Good" objects might be defined as the positive set, so that a positive sufficiency measure indicates that an object in more likely to be in the "Good" class, while a negative measure reflects that the object belongs to the "Bad" class. The sufficiency measure is now expressed as $M_S(x_{ki})$. A similar measure for necessity, $M_N(x_{ki})$, may be calculated assuming adherence to ϖ_1 corresponds to positive sufficiency and necessity measures.

$$M_N(x_{ki}) = M_N(x_{ki}, \varpi_1) \qquad \text{if } P(x_{ki} \mid \varpi_1) > P(x_{ki}) \qquad (18)$$

$$M_N(x_{ki}) = -M_N(x_{ki}, \varpi_2) \qquad \text{if } P(x_{ki} \mid \varpi_2) > P(x_{ki}) \qquad (19)$$

Hence it is possible to reduce the four AIM measures for the two classes into two measures, $M_S(x_{ki})$ and $M_N(x_{ki})$, where the sign of these measures is used to indicate to which class an object belongs.

Reducing the number of sufficiency and necessity measures from four to two eliminates the need for a summation scheme when calculating $M_V(x_{ki})$. Simple investigation of $M_S(x_{ki})$ and $M_N(x_{ki})$ indicates the relative importance of each attribute value.

The expression for the usefulness of an attribute as a whole is also simplified in the binary case. In cases where the necessity measure is not appropriate to the overall measure of the usefulness, as in Equation 9b, Equations 15 and 17 can be used to reduce $M_A(x_k)$ to the following simple summation.

$$M_A(x_k) = \sum_{i=1}^{n_k} M_S(x_{ki})^2 P(x_{ki}) \qquad (20)$$

8. Comparison with existing techniques

The publicly available "small" soybean dataset (Murphy & Aha, 1984) is used to compare the new AIM method presented here with existing classification and dimensionality reduction techniques. First used by Michalski (1980) and Michalski & Stepp (1983), the dataset contains 35 categorical attributes for 47 cases or instances. These instances are classified into four classes, (D1,D2,D3,D4), corresponding to different diseases in soybean plants. Each attribute has between one and seven values, and there are no missing values.

A detailed performance-based comparison between the AIM method and a large number of competing algorithms has not been conducted here. The main advantage of the AIM method is its ability to combine classification, dimensionality reduction and decision rule

extraction into a single method. Therefore, the purpose of this section is to demonstrate that this can be done with acceptable results.

Since the AIM method is most similar to a Naïve Bayes analysis, it is against this method that classification is compared. The "leave one out" cross validation error estimate (James, 1985, pp. 77-78, Hand et al. 2001, pp. 360) is used, due to the small sample size. Classification was carried out using a vector of the full 35 attributes, which were assumed to be independent. Using this method, both the Naïve Bayes and AIM methods misclassified 5 of the 47 cases, an error rate of 11%.

The AIM classification procedure produced measures of sufficiency, M_S, for each attribute value in each class. These are now summed, as in Equation 12, to provide a measure for the usefulness of each attribute, M_A. This is compared with the standard chi squared (χ^2) test, given in Equation 11, and the information gain method used in decision tree techniques (Quinlan, 1986). In all cases a high score indicates the attribute is useful in classification. The results are presented in Table 3. The attribute number given in the leftmost column corresponds to the order in which the data is provided in the literature (Murphy & Aha, 1984). The ranking of attributes based on their AIM, chi squared and information gain scores is also provided in the table. The results show that the three methods are in close agreement in their ranking of the bottom 21 attributes (that is, rank 15 and beyond). There is also parity in the identification of attributes 21 and 22 as being clearly the most useful. The correlation between the scores for the remaining attributes is also strong for the AIM and information gain methods.

Having ranked attributes in order of importance to classification, the data is re-classified using a reduced feature set. The number of attributes selected from Table 3 for this task may be found using cross validation, or similar methods. This process is not specific to the AIM method, and will therefore be omitted here. It is obvious from Table 3 that in this case attributes 21 and 22 are clearly superior and therefore they shall be the features selected for the re-classification of the data. The "leave one out" error rate is again used. The results of re-classification are now much improved, with 100% correct classification, using both the AIM method and Naïve Bayes. This demonstrates that the dimensionality reduction technique available through the AIM algorithm is viable, and is applicable to classification techniques other than the AIM method.

Table 3 Comparison of attribute scores and ranks using the AIM method (M_A), chi squared (χ^2) and information gain (Info gain) methods.

Attribute	M_A	χ^2	Info gain	M_A rank	χ^2 rank	Info gain rank
22	0.878	105.75	1.63	1	2	1
21	0.798	107.82	1.58	2	1	2
35	0.556	43.19	0.86	3	10	5
28	0.555	47.00	0.98	4	6	4
7	0.529	71.52	1.04	5	3	3
2	0.497	40.57	0.84	6	13	8
1	0.478	42.70	0.84	7	11	9
3	0.452	53.70	0.85	8	4	6
4	0.422	51.05	0.84	9	5	7
23	0.389	47.00	0.75	10	6	10
27	0.389	47.00	0.75	10	6	10
26	0.389	47.00	0.75	10	6	10
24	0.366	31.29	0.65	13	14	13
12	0.330	41.19	0.60	14	12	14
8	0.153	11.78	0.23	15	16	16
25	0.145	20.70	0.28	16	15	15
10	0.091	6.91	0.14	17	17	17
5	0.072	6.58	0.10	18	18	18
6	0.048	4.36	0.08	19	19	19
20	0.035	3.13	0.05	20	20	20
9	0.005	0.44	0.01	21	21	21
11	0	0	0	22	22	22
13	0	0	0	22	22	22
14	0	0	0	22	22	22
15	0	0	0	22	22	22
16	0	0	0	22	22	22
17	0	0	0	22	22	22
18	0	0	0	22	22	22
19	0	0	0	22	22	22
29	0	0	0	22	22	22
30	0	0	0	22	22	22
31	0	0	0	22	22	22
32	0	0	0	22	22	22
33	0	0	0	22	22	22
34	0	0	0	22	22	22

The AIM decision rule search algorithm is now applied to attributes 21 and 22 from the soybean data. This is a simple process, as most values for these attributes map directly to classes (hence their dominance in the attribute ranking). With a vector length of 1, $M_S(x_{ki}, \varpi_j) = 1$ for a number of attribute values, implying direct mapping, given below.

(Attribute 21 = 3) → D1 (21)
(Attribute 21 = 0) → D2 (22)
(Attribute 21 = 2) → D4 (23)
(Attribute 22 = 3) → D1 (24)
(Attribute 22 = 3) → D2 (25)
(Attribute 22 = 2) → D4 (26)

The necessity measure for the rules in Equations 22 and 25 are both 1, and they apply to the same class. Therefore, one of them maybe removed, as it is redundant to classification into D2. Also, the necessity measure for the rule in Equation 21 is 1, exceeding that of Equation 24, so that the latter may be removed. Equally, the necessity measure for Equation 26 is unity, greater than the measure for Equation 23, allowing the latter to be eliminated.

Class D3 is not represented in the rules above. Therefore, the vector length is increased to 2, and all attribute values satisfying $M_N(x_{ki}, D3) > 0$ are used to form new trial vectors. This applies to (Attribute 21 = 1) and (Attribute 22 = 1) only. Since $M_S((\text{Attribute } 21 = 1, \text{Attribute } 22 = 1), D3) = 1$, the following decision rule is obtained.

(Attribute 21 = 1, Attribute 22 = 1) → D3 (27)

Equations 21, 22, 26 and 27 provide a set of decision rules that are necessary and sufficient to accurately classify any new case, based on the training data.

9. Conclusion

A new method has been presented for classification of objects with categorical data. The Attribute Importance Measure (AIM) method uses logical conditions of sufficiency, necessity and irrelevance in combination with probabilistic techniques to provide features which are not available in traditional Naïve Bayes techniques. One such feature is the integration of dimensionality reduction into the classification algorithm.

The classification functionality of the algorithm has been shown to be comparable with Naïve Bayes, while the dimensionality reduction, or feature selection, gives similar results to a chi squared test. Indeed, the behaviour of the AIM attribute statistic for independent

data has equivalent properties to chi squared, such as independence from the probability density functions and a variation with the degrees of freedom.

It has been shown that the AIM method provides a mechanism by which searching for vectors of attribute values which are useful in classification (decision rules) may be performed efficiently.

Finally, the various aspects of the AIM method have been applied to an open source data set and compared with an existing classification and feature selection technique. Classification and dimensionality reduction performance was equivalent between the AIM method and existing techniques. This demonstrates the power of the AIM method to combine a number of tasks into one algorithm for real data.

10. References

Anderson, D.R., Sweeney, D.J., & Williams, T.A. (1994). Introduction to Statistics - Concepts and Applications, Third Edition, West Publishing Company.

Duda, R.O., Hart, P.E., Stork, D.G. (2001). Pattern Classification, Second Edition, Wiley-Interscience, New York.

Frank, E., Wang, Y., Inglis, S., Holmes, G., & Witten, I.H. (1998). Technical Note: Using model trees for classification, Machine Learning, 32, pp. 63-76.

Gennari, J.L., Langley, P., & Fisher, D. (1989). Models of incremental concept formulation, Artificial Intelligence 40, pp. 11-61.

James, M. (1985). Classification Algorithms, John Wiley & Sons, London, pp. 70-71.

Hand, D.J. (1981). Discrimination and classification, John Wiley & Sons.

Hand, D., Mannila, H., Smyth, P. (2001), Principles of Data Mining, MIT Press.

Hastie, T., Tibshirani, R., Friedman J. (2001). The Elements of Statistical Learning - Data Mining, Inference, and Prediction. Springer Series in Statistics. Springer.

Lewis, D.D. (1998). Naïve (Bayes) at Forty: The Independence Assumption in Information Retrieval, Proceedings of ECML-98, 10th European Conference on Machine Learning.

Lim, T., Loh, W., & Shih, Y. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms, Machine Learning, 40, pp. 203-228.

Michalski, R.S. (1980). Learning by being told and learning from examples: an experimental comparison of the two methods of knowledge acquisition in the context of developing an expert system for soybean disease diagnosis, International Journal of Policy Analysis and Information Systems, 4(2), 125-161.

Michalski, R.S., & Stepp, R.E. (1983). Automated construction of classifications: conceptual clustering versus numerical taxonomy, IEEE transactions on pattern analysis and machine intelligence, Vol. PAMI-5, No. 4, pp. 396-410.

Mosteller, F., & Wallace, D.L. (1984). Applied Bayesian and Classical Inference: The Case of The Federalist Papers, Springer-Verlag, New York.

Murphy, P.M., & Aha, D.W. (1994). UCI Repository of machine learning databases (http://www.ics.uci.edu/~mlearn/MLRepository.html), University of California, Department of Information and Computer Science.

Quinlan, J.R. (1986). Induction of Decision Trees, Machine Learning, Vol. 1, pp. 81-106.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization, ACM Computing Surveys, Vol. 34, No. 1, pp. 1-47.

11. Acknowledgements

Many thanks go to Dr Julian Sorensen, whose patient statistical advice expanded the scope of this paper while improving its mathematical rigour.

DISTRIBUTION LIST

A logical and probabilistic technique for classification and dimensionality reduction for objects with categorical data

Mark Porter

AUSTRALIA

DEFENCE ORGANISATION

Task Sponsor	
Scientific Adviser Joint	1
S&T Program	
Chief Defence Scientist	shared copy
FAS Science Policy	
AS Science Corporate Management	
Director General Science Policy Development	
Counsellor Defence Science, London	Doc Data Sheet
Counsellor Defence Science, Washington	Doc Data Sheet
Scientific Adviser to MRDC, Thailand	Doc Data Sheet
Navy Scientific Adviser	Doc Data Sht & Dist List
Scientific Adviser - Army	Doc Data Sht & Dist List
Air Force Scientific Adviser	Doc Data Sht & Dist List
Scientific Adviser to the DMO M&A	Doc Data Sht & Dist List
Scientific Adviser to the DMO ELL	Doc Data Sht & Dist List
Information Sciences Laboratory	
Chief of Intelligence, Surveillance and Reconnaissance Division	Doc Data Sht & Dist List
Research Leader - Secure Communications Branch	Doc Data Sht & Dist List
Ian Coat	1
Mark Porter	1
DSTO Library and Archives	
Library Edinburgh	2
Australian Archives	1
Capability Systems Division	
Director General Maritime Development	Doc Data Sheet
Director General Information Capability Development	Doc Data Sheet
Office of the Chief Information Officer	
Deputy CIO	Doc Data Sheet
Director General Information Policy and Plans	Doc Data Sheet
AS Information Strategies and Futures	Doc Data Sheet
AS Information Architecture and Management	Doc Data Sheet
Director General Australian Defence Simulation Office	Doc Data Sheet

Strategy Group

Director General Military Strategy — Doc Data Sheet
Director General Preparedness — Doc Data Sheet

HQAST

SO (Science) (ASJIC) — Doc Data Sheet

Navy

SO (SCIENCE), COMAUSNAVSURFGRP, NSW — Doc Data Sht & Dist List
Director General Navy Capability, Performance and Plans, Navy Headquarters — Doc Data Sheet
Director General Navy Strategic Policy and Futures, Navy Headquarters — Doc Data Sheet

Air Force

SO (Science) - Headquarters Air Combat Group, RAAF Base, Williamtown NSW 2314 — Doc Data Sht & Exec Summ

Army

ABCA National Standardisation Officer, Land Warfare Development Sector, Puckapunyal — e-mailed Doc Data Sheet
SO (Science), Deployable Joint Force Headquarters (DJFHQ) (L), Enoggera QLD — Doc Data Sheet
SO (Science) - Land Headquarters (LHQ), Victoria Barracks NSW — Doc Data Sht & Exec Summ

Intelligence Program

DGSTA Defence Intelligence Organisation — 1
Manager, Information Centre, Defence Intelligence Organisation — 1 printed & 1 pdf
Assistant Secretary Corporate, Defence Imagery and Geospatial Organisation — Doc Data Sheet

Defence Materiel Organisation

Head Aerospace Systems Division — Doc Data Sheet

Chief Joint Logistics Command — Doc Data Sheet
Head Materiel Finance — Doc Data Sheet

Defence Libraries

Library Manager, DLS-Canberra — Doc Data Sheet
Library Manager, DLS - Sydney West — Doc Data Sheet

OTHER ORGANISATIONS

National Library of Australia — 1
NASA (Canberra) — 1

UNIVERSITIES AND COLLEGES

Australian Defence Force Academy Library — 1

Head of Aerospace and Mechanical Engineering	1
Serials Section (M list), Deakin University Library, Geelong, VIC	1
Hargrave Library, Monash University	Doc Data Sheet
Librarian, Flinders University	1

OUTSIDE AUSTRALIA

INTERNATIONAL DEFENCE INFORMATION CENTRES

US Defense Technical Information Center	2
UK Defence Research Information Centre	2
Canada Defence Scientific Information Service	1
NZ Defence Information Centre	1

ABSTRACTING AND INFORMATION ORGANISATIONS

Library, Chemical Abstracts Reference Service	1
Engineering Societies Library, US	1
Materials Information, Cambridge Scientific Abstracts, US	1
Documents Librarian, The Center for Research Libraries, US	1

SPARES	5
Total number of copies: 30 printed copies	**31**

Page classification: UNCLASSIFIED

DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION
DOCUMENT CONTROL DATA

1. PRIVACY MARKING/CAVEAT (OF DOCUMENT)

2. TITLE

A Logical and Probabilistic Technique for Classification and Dimensionality Reduction for Objects with Categorical Data

3. SECURITY CLASSIFICATION (FOR UNCLASSIFIED REPORTS THAT ARE LIMITED RELEASE USE (L) NEXT TO DOCUMENT CLASSIFICATION)

Document	(U)
Title	(U)
Abstract	(U)

4. AUTHOR(S)

Mark Porter

5. CORPORATE AUTHOR

Information Sciences Laboratory
PO Box 1500
Edinburgh South Australia 5111 Australia

6a. DSTO NUMBER	6b. AR NUMBER	6c. TYPE OF REPORT	7. DOCUMENT DATE
DSTO-RR-0276	AR- 013-127	Research Report	June 2004

8. FILE NUMBER	9. TASK NUMBER	10. TASK SPONSOR	11. NO. OF PAGES	12. NO. OF REFERENCES
E8730/16/69	INT 03082	SA JOINT	27	16

13. URL on the World Wide Web

http://www.dsto.defence.gov.au/corporate/reports/DSTO-RR-0276.pdf

14. RELEASE AUTHORITY

Chief, Intelligence, Surveillance and Reconnaissance Division

15. SECONDARY RELEASE STATEMENT OF THIS DOCUMENT

Approved for public release

OVERSEAS ENQUIRIES OUTSIDE STATED LIMITATIONS SHOULD BE REFERRED THROUGH DOCUMENT EXCHANGE, PO BOX 1500, EDINBURGH, SA 5111

16. DELIBERATE ANNOUNCEMENT

No Limitations

17. CITATION IN OTHER DOCUMENTS Yes

18. DEFTEST DESCRIPTORS

classification, data mining, Bayes theorem and algorithms

19. ABSTRACT

A supervised learning technique, the Attribute Importance Measure (AIM) method, is proposed for the classification of objects with categorical attributes. The advantage of this method over existing techniques is its ability to perform classification and dimensionality reduction, or feature selection, with the same algorithm. The method uses probabilistic measures alongside logical concepts of sufficiency, necessity and irrelevance in providing corresponding weights to values in attribute value pairs. Finally an efficient search algorithm is developed which generates decision rules for classification. The performance of the new method is demonstrated on a commonly used machine learning data set.